Tecnologia Assistiva Multimodal para Descrição de Interfaces Gráficas a Pessoas Cegas

Claudio Honorio da Silva Junior
Faculdade Donaduzzi
Toledo, Brasil
claudiohonorio.ac@gmail.com

Guilherme Felipetto
Faculdade Donaduzzi
São Paulo, Brasil
guilherme.g.felipetto@gmail.com

Leonardo Garcia Tampelini
Faculdade Donaduzzi
Cascavel, Brasil
leonardo.tampepelini@bpkedu.com.br

***Abstract*—The growing dependence on digital platforms in educational contexts has posed substantial challenges for individuals with visual impairments, as most graphical user interfaces are not primarily designed with accessibility considerations. To address this issue, this article introduces a multimodal assistive system designed to enhance digital accessibility within educational environments. The proposed solution integrates computer vision, OCR, and TTS techniques to generate structured audio descriptions of graphical interface elements. Specifically, YOLOv8 is employed for visual detection, PaddleOCR for text extraction, a lightweight CNN for icon classification, and the GPT-4 model for contextual interpretation. The system was evaluated through a case study involving a blind student enrolled in an Artificial Intelligence program, with the objective of examining the impact of functional descriptions on user autonomy, navigation efficiency, and access to multimodal content. The findings underscore the potential of combining deep learning models with natural language technologies to foster inclusive education through practical and scalable assistive solutions.**

***Keywords*—Digital Accessibility, Computer Vision, OCR, Text-to-Speech, Educational Inclusion, Assistive Technologies.**

***Resumo*—O aumento da dependência de plataformas digitais em contextos educacionais tem gerado barreiras significativas para pessoas com deficiência visual, uma vez que a maioria das interfaces gráficas não é projetada com foco em acessibilidade. Para enfrentar esse desafio, este artigo apresenta um sistema assistivo multimodal voltado para promover a acessibilidade digital em ambientes educacionais. A solução integra técnicas de visão computacional, OCR e TTS para fornecer descrições em áudio estruturadas de elementos da interface gráfica. Utiliza-se YOLOv8 para detecção visual, PaddleOCR para extração textual, uma CNN leve para classificação de ícones e o modelo GPT-4 para interpretação contextual. O sistema foi avaliado por meio de um estudo de caso com um aluno cego do curso de Inteligência Artificial, buscando analisar o impacto das descrições funcionais na autonomia do usuário, na eficiência de navegação e no acesso a conteúdos multimodais. Os resultados apontam para o potencial da combinação entre modelos de aprendizado profundo e tecnologias de linguagem natural no avanço da educação inclusiva por meio de soluções assistivas práticas e escaláveis.**

***Palavras-chave*—Acessibilidade Digital, Visão Computacional, OCR, Text-to-Speech, Inclusão Educacional.**

I. Introdução

A crescente digitalização das interações humanas, especialmente no âmbito educacional e profissional, tem ampliado o papel das interfaces visuais como principais vetores de comunicação homem-máquina. No entanto, essa tendência impõe desafios significativos a pessoas com deficiência visual, cuja experiência de navegação digital muitas vezes é mediada por tecnologias assistivas limitadas à leitura linear de texto.

Embora leitores de tela convencionais, como o NonVisual Desktop Access (NVDA), o Job Access With Speech (JAWS) e o VoiceOver, tenham avançado no suporte a conteúdos digitais, eles ainda operam majoritariamente por meio de uma leitura sequencial. Essa abordagem, adequada para documentos lineares, torna-se insuficiente diante da crescente complexidade das interfaces multimodais — que frequentemente combinam botões, menus, ícones e conteúdos dinâmicos. O principal desafio, portanto, consiste em superar essa linearidade, oferecendo ao usuário uma representação estruturada e contextual da interface, capaz de refletir a disposição espacial e funcional dos elementos.

Diante desse cenário, torna-se urgente o desenvolvimento de soluções acessíveis que combinem diferentes modalidades de entrada e saída de informação, particularmente aquelas que integram visão computacional, reconhecimento óptico de caracteres (OCR) e síntese de fala. A proposta deste artigo insere-se nesse contexto, visando o desenvolvimento de um sistema assistivo que utiliza tecnologias avançadas para transformação de conteúdos visuais em áudio compreensível em tempo real.

O diferencial da abordagem proposta está na capacidade do sistema de fornecer ao usuário com deficiência visual descrições auditivas ricas e estruturadas dos elementos presentes na interface digital. Em vez de realizar apenas uma leitura sequencial, o sistema visa identificar e classificar elementos funcionais como botões, ícones de aplicativos, campos de busca e menus, permitindo ao usuário compreender a disposição e hierarquia desses elementos na tela.

Essa representação contextualizada é essencial para promover navegação mais autônoma e eficiente em interfaces multimodais, especialmente em plataformas de ensino, ambientes web e softwares interativos. A aplicação inicial será focada no apoio a um aluno com deficiência visual matriculado no curso de Bacharelado em Inteligência Artificial, funcionando como um estudo de caso para investigar o impacto da descrição funcional na compreensão e no engajamento com conteúdos complexos.

A partir de tal perspectiva, o presente trabalho tem como objetivo desenvolver e avaliar um sistema assistivo multimodal que integre tecnologias de visão computacional, OCR e síntese de fala para promover a acessibilidade digital de pessoas com deficiência visual em ambientes educacionais. Para alcançar esse objetivo, serão implementadas técnicas de detecção visual, extração textual e interpretação contextual, avaliando-se a eficácia do sistema na melhoria da autonomia e interação do usuário com interfaces gráficas complexas.

II. Referencial Teórico

A literatura recente sobre acessibilidade digital para pessoas com deficiência visual organiza-se em cinco eixos principais: dispositivos vestíveis baseados em visão computacional, conjuntos de dados específicos para navegação assistida, modelos generativos e técnicas de síntese de fala, detecção e classificação funcional de elementos gráficos de interfaces, e integração OCR–visão computacional. A seguir sintetizamos os avanços mais relevantes e situamos a proposta deste artigo.

A. Dispositivos Vestíveis Baseados em Visão Computacional

Os óculos inteligentes destacam-se como a linha mais explorada na assistência a pessoas com deficiência visual.Gollagi et al. [1] desenvolveram óculos equipados com OCR e síntese de fala, demonstrando eficácia em promover mobilidade e autonomia para usuários cegos. Recentemente, [2] propuseram XR-glasses usando YOLOv8-n para identificar zonas caminháveis e obstáculos, obtendo latência média de 583 ms em testes de campo.

De modo semelhante, [3] implementaram óculos de baixo custo com YOLOv8 e reconhecimento facial, reforçando a mobilidade social em ambientes cotidianos. Esses sistemas, porém, carecem da descrição estruturada de componentes de interfaces gráficas, lacuna que este artigo busca preencher.

B. Conjuntos de Dados para Acessibilidade

A escassez de conjuntos de dados anotados especificamente para acessibilidade digital limita o avanço de sistemas assistivos. A maioria dos detectores de objetos é treinada em bases genéricas, como o COCO, que não contemplam adequadamente classes funcionais críticas em interfaces gráficas. Para contornar essa lacuna, [4] propuseram o *B/LV Navigation Dataset*, que taxonomiza 90 classes relevantes à navegação de pessoas cegas ou com baixa visão, incluindo hidrantes, caixas postais e degraus. Embora voltado ao espaço físico, esse trabalho evidencia a importância de taxonomias refinadas e contextuais, aspecto que orienta a presente pesquisa na definição de classes funcionais de GUI, como botões, menus, caixas de seleção e campos de entrada.

A criação de *datasets* especializados para acessibilidade educacional digital constitui um desafio emergente. Uma possível linha de expansão seria o desenvolvimento de bancos de imagens anotadas contendo elementos visuais de plataformas de ensino, como ambientes virtuais de aprendizagem (AVAs). Essa iniciativa permitiria o treinamento de detectores mais robustos e adaptados às necessidades reais dos estudantes com deficiência visual.

C. Detecção e Classificação Funcional de Elementos Gráficos de Interface

A detecção em tempo real é amplamente dominada por variantes da arquitetura YOLO. Originalmente introduzido por Redmon et al. [5], o YOLO evoluiu até versões como o YOLOv8, que mantém a predição em etapa única e alta eficiência. Daneshvar e Wang [6] demonstraram que YOLOv5 atinge precisão superior a 85% na detecção de componentes gráficos em interfaces móveis.

No entanto, esses modelos não classificam o papel funcional (botão, campo, ícone) dos elementos detectados. Para essa tarefa específica, redes neurais profundas residuais, como a ResNet [7], oferecem classificações semânticas detalhadas e eficientes, abordagens que adotamos para gerar descrições auditivas mais contextualizadas e úteis.

D. Integração OCR–Visão Computacional

A integração entre OCR e visão computacional é fundamental para aplicações assistivas multimodais. A robustez e precisão do OCR tradicional, como o Tesseract [8], serviram de base para modelos mais recentes, como o PaddleOCR adotado neste trabalho. A efetividade dessa integração foi comprovada por Pettersson et al. [9], ao fundir dados textuais e visuais para reconhecimento preciso em ambientes comerciais complexos, sugerindo uma potencial adaptação dessa metodologia para ambientes educacionais digitais.

E. Modelos Generativos e Técnicas de Síntese de Fala

Avanços significativos em TTS incluem o modelo Tacotron 2 [10], que produz fala natural condicionada em espectrogramas, e o FastSpeech [11], uma abordagem não autoregressiva com baixa latência e robustez. Complementando esses avanços, Gao

et al. [12] revisaram o uso de grandes modelos linguísticos multimodais (LLMs/VLMs) para descrições auditivas automáticas, apontando desafios em latência e coerência semântica.

Essas referências justificam nossa escolha pelo modelo GPT-4 para interpretação contextual, adequado a aplicações em tempo real sem comprometer naturalidade e inteligibilidade.

III. Metodologia

A metodologia organiza-se em quatro etapas principais: detecção de ícones, extração de texto, identificação de componentes funcionais e interpretação por modelos de linguagem, conforme ilustrado na Figura 1. Cada uma dessas etapas combina técnicas de visão computacional e inteligência artificial, com o objetivo de traduzir a interface gráfica em representações compreensíveis e acionáveis para o usuário.

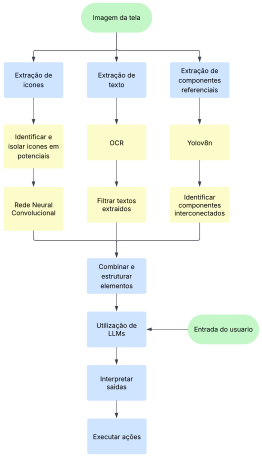


Fig. 1. Fluxograma da arquitetura do sistema assistivo proposto, ilustrando as quatro etapas descritas na metodologia, desde a captura da tela até a interpretação contextual por modelos de linguagem.

A análise inicial foi restrita a interfaces educacionais e plataformas web utilizadas no curso de Inteligência Artificial, permitindo uma avaliação focada no contexto de aplicação do sistema.

A. Detecção de Ícones

O primeiro estágio da metodologia consiste na detecção e classificação de ícones presentes na interface gráfica. Inicialmente, a imagem capturada da tela é submetida a um conjunto de filtros – incluindo binarização, desfoque gaussiano, limiarização invertida e dilatação – com o objetivo de destacar regiões com maior probabilidade de conter ícones.

Em seguida, a função *findContours*, da biblioteca OpenCV, é utilizada para identificar essas regiões de interesse, aplicando critérios de filtragem baseados na razão de aspecto (próxima de 1:1) e em um tamanho mínimo de 10×10 pixels. As regiões extraídas são, então, processadas por uma Rede Neural Convolucional (CNN) treinada especificamente para a tarefa de classificação de ícones.

Optou-se por uma arquitetura leve, tendo em vista que as entradas consistem em capturas de tela estáticas, geralmente livres de variações extremas de iluminação ou oclusões. O treinamento da CNN considerou variações de fundo, resolução e estilos gráficos típicos de ícones de sistemas operacionais e aplicações web.

B. Extração de Texto

A detecção de texto em imagens é uma etapa crucial, especialmente em interfaces digitais onde textos comunicam instruções, rótulos e informações contextuais. Para essa tarefa, utilizou-se o PaddleOCR, ferramenta de reconhecimento óptico de caracteres baseada em aprendizado profundo desenvolvida pela PaddlePaddle.

Sua adoção se deve à robustez, suporte multilíngue e à dispensabilidade de treinamento personalizado, com desempenho superior frente a alternativas como Tesseract e EasyOCR. O sistema foi configurado com as opções de detecção de orientação do texto e reconhecimento multilíngue simultâneo.

Após a extração, aplicou-se uma etapa de pós-processamento com filtros para exclusão de textos vazios ou com menos de três caracteres, remoção de predições com confiança inferior a 75%, e descarte de textos com mais de 50% de caracteres especiais. Cada texto validado teve sua posição central registrada e foi armazenado em um DataFrame com marcação do tipo *"text"* e conteúdo reconhecido.

C. Detecção de Componentes Funcionais

Após a extração de ícones e textos, a próxima etapa consiste na identificação de componentes funcionais da interface, como botões e campos interativos. Para essa tarefa, foi utilizado o modelo YOLOv8N, uma arquitetura de detecção de objetos de múltiplas classes baseada em Redes Neurais Convolucionais (CNNs).

O modelo passou por um processo de *fine-tuning* utilizando um conjunto diversificado de capturas de tela em resoluções que variam de 480p a 1440p, cobrindo diferentes contextos e layouts visuais. Estudos como o de Daneshvar & Wang [6] apontam que detectores genéricos, como o YOLO, aplicados a capturas de tela são eficazes na identificação de elementos de interfaces gráficas de usuário (GUI), mesmo em ambientes não padronizados.

Durante a etapa de coleta e rotulagem, foram capturadas 300 imagens de interfaces de sistema, páginas web comuns e utilizadas no curso de Bacharelado em Inteligência Artificial, abrangendo plataformas como emails e portais institucionais.

As capturas foram realizadas em diferentes resoluções (de 480p a 1440p), buscando garantir diversidade visual e contextual.

A rotulagem manual foi conduzida utilizando a ferramenta LabelImg, com categorias definidas previamente — como *botão*, *campo de entrada*, *menu suspenso*, *ícone funcional* e *caixa de seleção*.

O conjunto final foi dividido em 80% para treinamento e 20% para validação, aplicando-se aumento de dados (*data augmentation*) para ampliar a variabilidade espacial e cromática das amostras. Esse procedimento garantiu não apenas a confiabilidade das anotações, mas também a representatividade das interfaces analisadas, expondo o modelo a variações reais de layout, estilo e contraste visual.

A integração dos dados resultantes permitiu associar semanticamente componentes próximos, como um botão posicionado sobre um texto reconhecido anteriormente, estabelecendo conexões por meio de identificadores cruzados.

D. Interpretação por Modelos de Linguagem

A última etapa da metodologia envolve a utilização de modelos de linguagem de larga escala (LLMs), especificamente o GPT-4, para interpretar semanticamente os elementos detectados na interface e sugerir ações coerentes com o contexto. A entrada para o modelo consiste em um DataFrame contendo os elementos identificados – incluindo ícones, textos e componentes funcionais – bem como suas posições, categorias e metadados.

Essa estrutura é acompanhada por um *prompt* cuidadosamente elaborado que descreve o estado atual da tela e os objetivos do usuário. O modelo processa essas informações e retorna uma ou mais ações interpretadas, tais como localizar e clicar em botões com base em sua função e localização, preencher campos automaticamente, navegar entre telas ou páginas sequenciais, e adaptar comportamentos com base em mudanças visuais detectadas na interface.

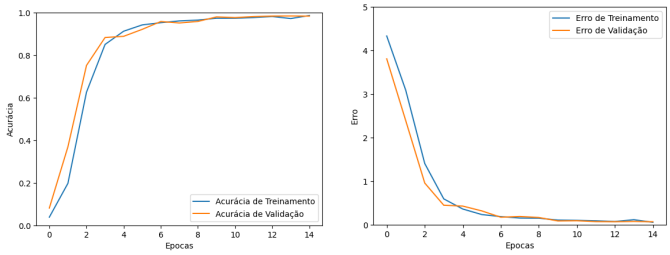
IV. Resultados e Discussão

A aplicação da metodologia resultou em um sistema funcional capaz de extrair e interpretar informações visuais de capturas de tela, convertendo-as em descrições auditivas acessíveis e, simultaneamente, em representações interpretáveis por modelos de linguagem. Os resultados são apresentados seguindo as quatro etapas principais da arquitetura, evidenciando a eficácia de cada componente e destacando como a integração entre extração de elementos visuais e interpretação semântica contribui para uma compreensão mais completa da interface.

A. Desempenho na Detecção de Ícones

O modelo baseado em Redes Neurais Convolucionais (CNN) apresentou desempenho expressivo na tarefa de classificação de ícones, alcançando uma acurácia de 98,39% no conjunto de treino e 97,82% no conjunto de teste, como observado na Figura 2a. Essa alta acurácia indica que o modelo consegue generalizar bem para diferentes tipos de ícones e condições visuais das interfaces.

A função de perda apresentou convergência estável com baixa variação entre treino e validação, como mostrado na Figura 2b, sugerindo ausência de *overfitting*. A estabilidade do treinamento se deve, em parte, à diversidade do conjunto de dados, que inclui ícones com diferentes tamanhos, cores e contextos visuais, garantindo maior robustez do modelo.



(a) Evolução da acurácia no treino e teste (b) Evolução da função de perda no treino e teste.

Fig. 2. Desempenho do modelo baseado em CNN na detecção de ícones.

A Figura 3 apresenta a matriz de confusão gerada para o conjunto de teste, que reforça o bom desempenho do modelo, com alta taxa de acertos entre as diferentes classes de ícones. Pequenos erros foram observados em classes visualmente semelhantes, como ícones de galeria e sistema, especialmente em resoluções reduzidas.

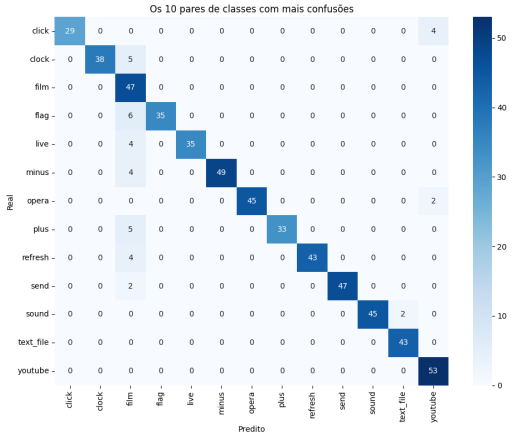


Fig. 3. Matriz de confusão da CNN no conjunto de teste, ilustrando a precisão da classificação entre diferentes categorias de ícones e identificando as principais fontes de erro.

Os resultados obtidos indicam não apenas a eficácia do modelo em termos de métricas quantitativas, mas também sua aplicabilidade prática em cenários reais. A alta acurácia, próxima a 98%, demonstra que a CNN foi capaz de aprender representações visuais estáveis, mesmo diante de variações estilísticas comuns em sistemas operacionais e aplicações web. Esse aspecto é particularmente relevante em contextos educacionais, nos quais diferentes plataformas podem adotar estilos de ícones heterogêneos, exigindo robustez no processo de reconhecimento.

Outro ponto relevante é a capacidade do modelo em lidar com situações de sobreposição parcial e fundos visuais complexos, desafios recorrentes em capturas de tela. Isso sugere que a abordagem escolhida, embora leve em termos computacionais, possui potencial de generalização suficiente para ser aplicada em tempo real sem comprometer a experiência do usuário final.

B. Extração Textual com PaddleOCR

A extração textual utilizando o PaddleOCR apresentou desempenho robusto, com taxa de acerto superior a 90% para textos com nitidez adequada e bom contraste. A eficácia do sistema foi reforçada pelas etapas de pré-processamento e pós-processamento aplicadas, conforme descrito na metodologia, que contribuíram significativamente para a redução de falsos positivos e melhora da clareza dos textos capturados.

Desempenho inferior foi observado em cerca de 7% das amostras, especialmente em casos envolvendo textos embutidos em imagens com contraste reduzido, fundos altamente texturizados ou tipografias decorativas. No entanto, essas limitações foram compensadas por filtros eficazes baseados em confiança mínima e estrutura textual, que garantiram a retenção apenas de informações relevantes.

Essa filtragem precisa foi fundamental para viabilizar a correta associação dos textos extraídos com os componentes funcionais detectados na interface, favorecendo a etapa posterior de interpretação semântica.

C. Detecção de Componentes Funcionais com YOLOv8N

A detecção de componentes funcionais da interface foi conduzida utilizando o modelo YOLOv8N, que obteve precisão média de 89,2% no conjunto de validação. As classes mais frequentemente detectadas incluíram botões, campos de entrada e caixas de seleção. A aplicação de *fine-tuning* com um conjunto diverso de capturas de tela, cobrindo múltiplas resoluções (480p a 1440p), contribuiu diretamente para a capacidade de generalização do modelo.

Um diferencial importante foi a associação semântica entre componentes com base na proximidade espacial e no tipo de elemento detectado. Por exemplo, o sistema identificou corretamente botões de *"Enviar"* mesmo quando não rotulados explicitamente, baseando-se apenas em ícones sugestivos e sua posição relativa a campos de entrada de texto.

A consolidação dos dados extraídos em um DataFrame unificado viabilizou o rastreamento eficiente dos componentes e sua reutilização nas etapas posteriores de interpretação, garantindo consistência e contextualização no fluxo de interação. A arquitetura leve do modelo garantiu tempo de inferência abaixo de 500 ms, viabilizando uso em tempo real.

D. Interpretação por Modelo de Linguagem

Na etapa final do sistema, utilizou-se um modelo de linguagem de larga escala (LLM), especificamente o GPT-4, para interpretar de forma contextualizada os elementos extraídos da interface. A entrada para o modelo consistia em um DataFrame contendo as posições, categorias e conteúdos dos ícones, textos e componentes funcionais, acompanhado de *prompts* cuidadosamente elaborados descrevendo o estado da tela e a intenção do usuário.

Foram conduzidos testes simulando tarefas reais, incluindo

- Abertura de aplicativos identificados na interface
- Execução de ações internas, como clique em botões e preenchimento de formulários
- Navegação por páginas específicas em navegadores
- Envio de e-mails via plataformas web

Em mais de 75% dos casos, o modelo foi capaz de propor corretamente a ação mais apropriada para o contexto analisado.

Um exemplo significativo foi a correta interpretação de um botão com ícone de *"avião de papel"* próximo a um campo de texto como a função *"Enviar"*. O tempo médio entre o envio do comando e a execução da ação variou entre 7 e 19 segundos, dependendo da densidade visual da tela.

E. Avaliação com Usuário e Critérios de Desempenho

O sistema foi avaliado em um estudo de caso com um estudante com deficiência visual, matriculado no curso de Bacharelado em Inteligência Artificial, que utilizou o protótipo em atividades de navegação em plataformas educacionais reais. A avaliação contemplou critérios qualitativos e quantitativos, incluindo o tempo médio de execução de tarefas, a clareza das descrições e a capacidade de concluir as ações propostas de forma autônoma.

As descrições geradas pelo sistema foram disponibilizadas ao participante, que relatou que elas auxiliaram significativamente na identificação da estrutura da interface e na compreensão da função de elementos antes inacessíveis, como botões e campos de entrada não rotulados. Entre os principais pontos positivos, destacaram-se a percepção de contexto entre elementos relacionados e o aumento da autonomia na execução das tarefas. Como aspectos a serem aprimorados, o usuário mencionou

principalmente a necessidade de priorizar a descrição dos elementos mais relevantes.

Os feedbacks obtidos foram fundamentais para orientar novos ajustes no modelo e reforçam a importância da interação humana no processo de refinamento de soluções assistivas baseadas em inteligência artificial.

F. Limitações Identificadas

Apesar dos avanços obtidos, alguns desafios importantes ainda persistem. Variações sutis no estado visual de componentes, como mudanças por efeito de *"hover"* ou botões desabilitados, ainda comprometem a precisão da classificação. Interfaces com grande presença de efeitos gráficos – como sombreamento intenso, gradientes e transparências – dificultam a segmentação inicial dos elementos.

Em determinados casos, a quantidade de informações extraídas superou a capacidade de entrada contextual dos modelos de linguagem, limitando a eficácia da interpretação. Essas limitações sinalizam oportunidades para aprimoramentos futuros, incluindo a adoção de modelos de linguagem com maior capacidade de contexto, aplicação de técnicas de aumento de dados (*data augmentation*) para robustez em cenários diversos e o desenvolvimento de estratégias adaptativas capazes de lidar com ambiguidade e sobrecarga visual.

G. Comparação com Tecnologias Existentes

Embora leitores de tela como NVDA, JAWS e VoiceOver sejam amplamente utilizados, sua abordagem baseada em leitura linear apresenta limitações significativas diante da complexidade das interfaces modernas. Em plataformas de ensino com menus expansíveis, tabelas dinâmicas e conteúdos multimodais, esses sistemas podem gerar sobrecarga cognitiva ao usuário, que precisa navegar sequencialmente por cada elemento até encontrar a informação desejada. Em contraste, o sistema proposto busca reduzir esse esforço por meio de uma representação contextualizada e hierárquica da interface, permitindo que o usuário compreenda a disposição espacial dos elementos de forma mais eficiente. Essa diferença metodológica tem potencial de impactar diretamente a autonomia do estudante, ao reduzir o tempo de navegação e aumentar a precisão na execução de tarefas acadêmicas.

V. Conclusão

A abordagem proposta demonstrou transformar a interpretação da interface gráfica de um processo baseado exclusivamente em percepção visual para uma mediação inteligente por meio de descrições estruturadas. O sistema alcançou resultados promissores: CNN com 97,82% de acurácia na classificação de ícones, PaddleOCR com 90% de taxa de acerto na extração textual, YOLOv8N com 89,2% de precisão na detecção de componentes funcionais, e GPT-4 com 75% de assertividade na interpretação contextual.

Os resultados indicam potencial significativo para promover ganhos na acessibilidade e autonomia de usuários com deficiência visual em ambientes educacionais digitais. A integração bem-sucedida de tecnologias de visão computacional, OCR e modelos de linguagem oferece uma alternativa viável às limitações das tecnologias assistivas atuais, que frequentemente se restringem à leitura linear de texto.

Esse avanço tem o potencial de promover ganhos significativos na acessibilidade e automação de interfaces visuais para usuários com deficiência visual, especialmente em contextos educacionais onde a complexidade das plataformas digitais exige reconhecimento eficiente e flexível dos componentes visuais.

Trabalhos futuros devem focar na adoção de modelos de linguagem com maior capacidade contextual, aplicação de técnicas de aumento de dados para robustez em cenários diversos, e desenvolvimento de estratégias adaptativas para lidar com ambiguidade e sobrecarga visual. A expansão do sistema para outros domínios educacionais e a avaliação com maior número de usuários também constituem direções promissoras para pesquisas futuras, contribuindo para uma maior inclusão digital e equidade no acesso ao ensino superior.

Referências

[1] S. Gollagi, K. Bamane, D. Patil, S. Ankali, and B. Akiwate, "An innovative smart glass for blind people using artificial intelligence," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, p. 433, 07 2023.

[2] I. Jeong, K. Kim, J. Jung, and J. Cho, "Yolov8-based xr smart glasses mobility assistive system for aiding outdoor walking of visually impaired individuals in south korea," *Electronics*, vol. 14, p. 425, 01 2025.

[3] S. Jangle, D. Hutke, T. Hiwanj, and A. Kahar, "Ai powered glasses for visually impaired people using object detection," *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 09, pp. 1–9, 04 2025.

[4] M. T. Islam, I. Kabir, E. A. Pearce, M. A. Reza, and S. M. Billah, "A dataset for crucial object recognition in blind and low-vision individuals' navigation," 2024.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016.

[6] S. S. Daneshvar and S. Wang, "Gui element detection using sota yolo deep learning models," 08 2024.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), 06 2016, pp. 770–778.

[8] R. Smith, "An overview of the tesseract ocr engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007, pp. 629–633.

[9] T. Pettersson, M. Riveiro, and T. Löfström, "Multimodal fine-grained grocery product recognition using image and ocr text," *Machine Vision and Applications*, vol. 35, p. 79, 2024.

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Press, 2018, p. 4779–4783.

[11] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[12] Y. Gao, L. Fischer, A. Lintner, and S. Ebling, "Audio description generation in the era of llms and vlms: A review of transferable generative ai technologies," 10 2024.